

***FORMATION DES CHERCHEURS IMPLIQUES
DANS LES PROJETS IRAD/C2D***

Sur le thème :

**« PRATIQUE DES TECHNIQUES D'ANALYSES FACTORIELLES
MULTIVARIEES, AVEC LE LOGICIEL R, DE DONNEES ISSUES
D'ENQUETES ET D'EXPERIMENTATIONS AGRICOLES »**

Yaoundé, 09 –13 septembre 2013

RAPPORT DES FORMATEURS



Par :

Dr. Christian CILAS
Cirad, Montpellier - France
et

Dr. Michel NDOUMBE NKENG
IRAD, Yaoundé- Cameroun

Septembre 2013

SOMMAIRE

1. Introduction	3
2. Objectifs de l'atelier	3
3. Déroulement de l'atelier	3
4. Résultats obtenus	5
5. Conclusion et recommandations	6
Annexes	7
Annexe 1 : Liste des participants	
Annexe 2 : Fiche de présence	
Annexe 3 : Chronogramme détaillé de l'atelier	
Annexe 4 : Comptes rendus journaliers (des participants)	

1. INTRODUCTION

Dans le but d'assurer une mise à niveau des chercheurs impliqués dans les projets IRAD/C2D en matière de maîtrise d'utilisation de logiciel de traitement et d'analyses statistiques des données, l'équipe de coordination des projets « C2D-PAR » a organisé un premier atelier du 3 au 7 juin 2013. Le thème de cet atelier portait sur « les techniques de collecte et de traitement statistique, avec le logiciel R, des données issues d'expérimentations agricoles ». Suite à cet atelier, une des recommandations formulées était l'organisation d'une autre session de formation, portant sur des modules statistiques plus avancés. C'est dans ce cadre que s'inscrit cette deuxième session de formation. Outre la présence du Biostatisticien IRAD, les participants ont également bénéficié de la présence d'un Biostatisticien senior du CIRAD, avec qui l'atelier a été co-animé.

Le présent rapport fait état du déroulement des travaux et de quelques recommandations formulées à l'issue de cette seconde session de formation.

2. OBJECTIFS DE LA FORMATION

L'objectif général de la formation était de familiariser les chercheurs à la pratique des techniques d'analyses factorielles multivariées afin de mieux valoriser les données issues d'enquêtes ou d'expérimentations agricoles.

.

De manière spécifique, il s'est agi de :

- Parcourir les principes généraux d'expérimentation agricole et d'enquête;
- Effectuer et interpréter quelques techniques classiques d'analyses statistiques;
- Familiariser les chercheurs à l'utilisation d'un logiciel spécialisé d'analyses statistiques, R.

3. DEROULEMENT DE L'ATELIER

L'atelier s'est déroulé durant cinq jours consécutifs, du 9 au 13 septembre 2013. Les horaires de travail allaient de 9h00 à 17h00. Chacune des journées était ponctuée par deux



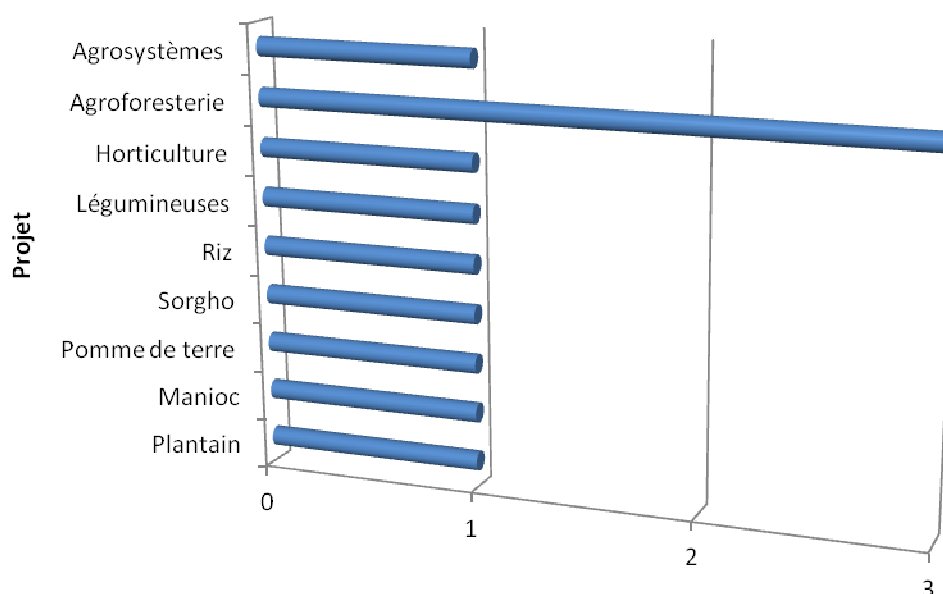
pauses (10h30 – 11h00, 13h00-14h00). La première journée de la formation, co-présidée par le Directeur Régional du CIRAD et le Directeur Général Adjoint de l'IRAD, a commencé par un partage des termes de référence. Cet exercice s'est effectué en deux phases successives : (i) présentation de l'orientation et contenu de l'atelier et (ii) présentation de la démarche pédagogique. A partir du deuxième jour, les travaux commençaient par une restitution de la journée précédente, faite par les participants.

a. Les participants

Il y a eu 12 participants, représentant les dix projets C2D-PAR, notamment les projets Plantain, Manioc, Pomme de terre, Sorgho, Riz, Légumineuses, Horticulture, Agroforesterie, Volaille, et Agrosystèmes (Figure 1), avec 25% des participants du genre féminin. La formation a été assurée par Dr. Christian CILAS (Biométricien – CIRAD) et Dr. Michel NDOUMBE NKENG, Biométricien – IRAD. (Voir liste des participants en annexe 1).



Figure 1: Distribution des participants par projet



b. Démarche pédagogique

Pendant la formation, les consultants ont alterné exposés théoriques et travaux pratiques sur le logiciel R. Au cours des travaux pratiques, les participants ont travaillé chacun sur leur poste (ordinateur portable). Les participants ont été responsabilisés par projet pour prendre des notes afin d'élaborer le compte rendu présenté le lendemain ; ils assuraient également le respect du timing. Les exposés des restitutions journalières ont servi aux formateurs de moyen d'évaluation progressive du degré de maîtrise des connaissances acquises. Les divers exposés des formateurs ont été remis aux participants.



Une évaluation a sanctionné la fin de la formation. Elle a consisté à emmener les « formés » à porter un jugement sur la pertinence des thèmes, la démarche pédagogique et le degré d'assimilation. En fin de formation, des attestations ont été remises aux participants.

4. RESULTATS OBTENUS

À l'issue de cet atelier, les participants sont capables de :

- Effectuer et interpréter une analyse en composantes principales (*Principal Component Analysis*);
- Effectuer et interpréter une analyse des correspondances (*Correspondance analysis*) ;
- Effectuer et interpréter une analyse discriminante (*Discriminant analysis*)
- Effectuer et interpréter une analyse de classification ascendante hiérarchique (*Cluster analysis*).

Les participants se sont également davantage familiarisés à l'utilisation du logiciel d'analyses statistiques R, logiciel statistique libre et gratuit.

5. CONCLUSION ET RECOMMANDATIONS

Il est possible d'affirmer au vu des résultats de l'évaluation par les participants, que les travaux de cet atelier se sont bien déroulés ; l'ambiance était cordiale et détendue ; les objectifs liés à cette formation ont été atteints. Les participants ont globalement trouvé l'environnement de travail calme et convivial. Les méthodes pédagogiques et le contenu de la formation ont été jugés très satisfaisants.

Certains participants ont néanmoins eu à déplorer le cadre de travail qu'ils ont trouvé restreint, avec une connexion Internet pas toujours satisfaisante. D'autres participants ont estimé que, lors des exposés théoriques, les formateurs étaient par moments très rapides. Pour d'autres encore, la partie du cours sur l'analyse spatiale leur a paru superflue.

Les formateurs ont exprimé leur disponibilité à répondre aux sollicitations éventuelles des participants.

A l'issue de cet atelier, la principale recommandation formulée est que les participants présentent à leurs équipes respectives et aux autres membres des projets ce qu'ils ont appris durant ces sessions de formation. Ceci les amènerait, non seulement à revisiter de fond en comble les notions acquises, mais aussi à partager ces connaissances avec leurs autres collègues.

ANNEXES

Annexe 1

FORMATION DES CHERCHEURS IMPLIQUES DANS LES PROJETS IRAD/C2D

Yaoundé, 09 – 13 septembre 2013

LISTE DES PARTICIPANTS

N°	NOM ET PRENOM	PROJET C2D	PROVENANCE	ADRESSE	
				Téléphone	E-mail
1	AMANG A MBANG Joseph	Agroforesterie (SAF)	Nkolbisson	77 39 08 49	mbang4@yahoo.com
2	BINDZI Boniface	Riz	Nkolbisson	96 04 79 64	bindzi_boniface@yahoo.fr
3	CILAS Christian	Formateur	Montpellier		cilas@cirad.fr
4	DJONNEWA André	Sorgho	Nkolbisson	75 74 52 22	djonnewa55@yahoo.com
5	EKALLE NYAME Murielle C.	Volailles	Nkolbisson	96 30 60 28	ekallemurielle@yahoo.fr
6	ETAME Francis	Manioc	Nkolbisson	76 29 84 22	etamefrancis@yahoo.fr
7	MBONGO LOMBI Franklin	Plantain	Njombé	77 53 36 09	fmbongo2001@yahoo.fr
8	MBOUSSI A MESSIA	Légumineuses	Ekona	75 84 26 91	mboussiemma@yahoo.com
9	MEKA SINDJE SOLANGE épouse DIEUMOU	Pomme de terre	Bambui	78 12 37 77	meka_sol@yahoo.fr
10	NDO Eunice épouse MFOU'OU	Horticulture	Nkolbisson	94 81 36 60	ndo_eunice@yahoo.fr
11	NDOUMBE NKENG Michel	Formateur	Douala	77 40 13 94	ndoumbe.nkeng@yahoo.com
12	OLINA BASSALA Jean Paul	Agro-système Nord	Garoua	77 47 49 46	olina_jp@hotmail.fr
13	PAGANI Vérane	Chargé de mission C2D	Yaoundé	90 64 52 13	verane.pagani@cirad.fr
14	TADU Zéphirin	Agroforesterie	Yaoundé	77 60 21 24	ztadu@yahoo.fr
15	TODEM NGNOGUE Hervé	Agroforesterie	Yaoundé	75 70 67 91	th.ngnogue@yahoo.fr

FORMATION DES CHERCHEURS IMPLIQUES DANS LES PROJETS IRAD/C2D

Yaoundé, 09 – 13 septembre 2013

FICHE DE PRESENCE

N°	NOM ET PRENOM	Date				
		09.09.2013	10.09.2013	11.09.2013	12.09.2013	13.09.2013
1	EKALIE NYAME Murielle					
2	TODEN NGONGUE Hesvi					
3	NDO Eunko					
4	ETAME Francis					
5	Meke Calange					
6	Amang a Mbong Joseph					
7	TADU ZEMIRIN					
8	MBONGO LOMBI Franklin					
9	JDONNENA Andre					
10	Prindzi Boniface					
11	MBOUSSI A MESSIA					
12	OLINA BASSALA Jean Paul					
13	MBONGUE NGONGUE					
14	ETAME					
15						

ANNEXE 3 :

CHRONOGRAMME DETAILLE DE L'ATELIER (*Yaoundé, 9 – 13 septembre 2013*)

Horaires	Journée				
	09.09.2013	10.09.2013	11.09.2013	12.09.2013	13.09.2013
9H00 -10H30	<ul style="list-style-type: none"> Inscription des participants Partage des TDR Désignation des rapporteurs Mot DR/CIRAD Mot d'ouverture (DG/IRAD) 	<ul style="list-style-type: none"> Restitution jour 1 Analyse bivariee: Analyse des corrélations 	<ul style="list-style-type: none"> Restitution jour 2 <p>Test d'indépendance de deux caractères qualitatifs Analyse factorielle des correspondances (AFC)</p>	<ul style="list-style-type: none"> Restitution jour 3 <p>Classification Ascendante Hiérarchique (CAH)</p>	<ul style="list-style-type: none"> Restitution jour 4 <p>Introduction aux techniques d'analyse spatiale (Notions de géostatistique)</p>
10H30 – 11H00	<i>Pause café</i>	<i>Pause café</i>	<i>Pause café</i>	<i>Pause café</i>	<i>Pause café</i>
11H00 – 13H00	Rappel des notions de Statistique descriptive	Analyse en composantes principales (ACP)	<u>Exercices pratiques</u> Analyse factorielle des correspondances (AFC)	<u>Exercices pratiques</u> Classification Ascendante Hiérarchique (CAH)	<ul style="list-style-type: none"> Synthèse formation Evaluation de l'atelier Fin de l'atelier
13H00 – 14H00	<i>Pause déjeuner</i>	<i>Pause déjeuner</i>	<i>Pause déjeuner</i>	<i>Pause déjeuner</i>	<i>Pause déjeuner</i>
14H00 – 17H00	<u>Exercices pratiques</u> Rappel Statistique descriptive : techniques graphiques et résumés numériques	<u>Exercices pratiques</u> Analyse en composantes principales (ACP)	<u>Exercices pratiques</u> Analyse factorielle des correspondances multiples (ACM)	Introduction à l'analyse discriminante (AD)	

NB : *Tous les exercices pratiques se feront avec le logiciel R*

Annexe 4 : COMPTES RENDUS JOURNALIERS

COMPTE RENDU DU 09 SEPTEMBRE 2013

Rapporteurs : Djonnewa André et Meka Solange

Introduction

L'atelier sur la biométrie à l'attention des équipes des projets C2D-PAR a commencé le 09 septembre 2013 à 9h 20 mn au CIRAD par la présentation des participants suivi de celle du programme par Dr Michel Ndoumbé.

.cette étape a laissé place au mot du Délégué Régional du CIRAD qui a insisté sur l'importance de cette formation qui vient à point nommé au moment de la mise en place des projets C2D-PAR. Le Délégué Régional du CIRAD a signalé qu'il y aura des missions de suivi tout au long de l'exécution des projets, d'où l'organisation de cette formation dispensée par des formateurs de haut niveau. Il a seulement regretté que tous les projets ne soient pas représentés. Cependant, il a rassuré qu'il allait voir avec la Direction générale de l'IRAD quelle disposition prendre pour que tous les représentants des projets puissent assister à cet atelier au cours de la semaine.

Après le mot du Directeur régional du CIRAD, Dr. Christian Cilas a fait une présentation sur l'initiation aux statistiques.

Cette première phase de l'atelier de formation s'est appesantie sur deux points essentiels à savoir : la définition de la statistique et les éléments de terminologie.

1. Définition de la statistique

Le terme statistique, venant du Latin *status* = « état », a été définie comme une étude qui consiste à observer et à étudier une *particularité commune* chez un *groupe de personnes ou de choses*. Les notes d'un groupe d'élèves ou le taux de mortalité par cancer des fumeurs par exemple.

Il existe deux domaines en statistique :

La Statistique descriptive (exploration des données, mesures et indices, représentations graphiques) et la Statistique inférentielle (tester des hypothèses, faire des prédictions à partir d'échantillons). Pour expliciter ces notions, des exemples ont été donnés.

(1) Lorsqu'on calcule la moyenne d'une classe par exemple, on fait de la statistique descriptive, mais quand on dit la longueur des phrases dans le Petit Prince est significativement plus courte que dans le journal Le Monde (25 mots), il s'agit de la statistique inférentielle.

2. Quelques éléments de terminologie

Le formateur a rappelé qu'une étude statistique consiste à observer et à étudier une particularité commune chez un groupe de personnes ou de choses. Ce rappel a introduit la définition de quelques termes comme :

Population qui est un ensemble de tous les objets que l'on étudie (des personnes, un groupe d'élèves, un groupe de malades, etc.)

Individus qui sont les éléments de la population étudiée (chacune des personnes interrogées pour une enquête, chaque jour de l'année pour lequel on dispose de données météorologiques, etc)

Variable ou attribut ou caractère qui est une propriété commune aux individus de la population que l'on souhaite étudier (la taille de personnes, la couleur de leurs yeux, leur sexe etc.)

Il existe deux types de variables, elles peuvent être :

Qualitatives (sexe, profession) ou

Quantitatives (taille, poids, âge, taux de cholestérol, etc)

Modalités : les modalités d'une variable qualitative sont les différentes valeurs que celle-ci peut prendre (variable **situation familiale** : *célibataire, marié, veuf, divorcé*, variable **sexe** : *homme, femme*).

Les Variables quantitatives sont de deux types :

Les variables discrètes elles prennent des valeurs dénombrables (ex.: nombres entiers) c'est l'exemple de nombre d'enfants par famille ou de nombre de pièces d'un appartement

Les variables continues qui prennent une infinité de valeurs, aussi proches qu'on veut (ex.: nombres réels) comme la taille ou le poids des personnes

La mesure ou le dénombrement

Une différence a été notée entre ces deux termes en ce sens qu'on dénombre les individus dans telle ou telle modalité d'une (65 personnes aux yeux bleus) alors qu'on mesure les variables quantitatives (taille de l'individu X = 1,75 m)

A 10h30 mn, la pause est annoncée au moment où arrive le Directeur Général Adjoint de l'IRAD qui déclare n'avoir rien de particulier à signaler sinon inviter les participants à bien suivre la formation.

Les travaux ont repris à 11h 05 mn et ont été consacrés aux méthodes statistiques telles :

Les méthodes descriptives qui résument l'information graphique de distribution, paramètres statistiques et les analyses multivariées (ACP, AFC, etc.)

Les méthodes inférentielles qui ont pour objectifs de tester les hypothèses et de comparer les traitements et les outils tels que les dispositifs expérimentaux et les analyses de variance tests

Les méthodes descriptives ont été illustrées par des graphiques de distribution tel que l'histogramme, représentation de la distribution d'une variable quantitative (exemple : numero de branches en production par tige de caféier), le Stem and leaf (tige et feuille), le box plot (boîte de dispersion ou boîte à moustache) qui permet de comparer les distributions et de déterminer le premier quartile : $Q0.25$, le second quartile : $Q0.75$ et l'intervalle inter quartiles : $IQR = Q0.75 - Q0.25$. Des paramètres statistiques qui permettent de décrire les distributions ont été également mentionnés, notamment les paramètres de centralité (moyenne, médiane, mode), les paramètres de dispersion (amplitude, variance, écart-type) ainsi que les paramètres de symétrie (Pearson, Fisher (loi normale: 0)) et les paramètres d'aplatissement (Pearson, Fisher (loi normale: 3 ; 0)). Les lois de probabilité ont été tout de même enseignées. Parmi elles, on distingue, pour les variables variables discrètes, la loi binomiale, de Poisson et la loi binomiale négative, pour les variables continues la loi normale, la loi lognormale et les lois dérivées des lois normales (χ^2 , Student et Fisher).

Des exercices d'illustration ont occupé l'après midi de la première journée qui s'est achevée par l'installation du logiciel Package R.

COMPTE RENDU DU 10 SEPTEMBRE 2013

Rapporteurs : Mboussi à Messia et Bindzi Boniface

Plan

- Restitution compte rendu du premier jour
- Rappels théoriques sur la corrélation et la régression
- Exercice pratique sur la corrélation/ régression
- L'analyse de la composante principale
- Exercice pratique

Compte rendu du 09/09/2013

- Après la restitution du compte rendu du groupe 1 par nos collègues des C2D sorgho et pomme de terre, les formateurs à savoir les Drs Ndoumbe et Cilas ont apporté des compléments et corrections pour compléter le rapport.

Rappels théoriques sur la corrélation et la régression

- La corrélation(lien linéaire entre 2 variables) étudie la relation entre 2 ou plusieurs variables sans se soucier de leur causalité. Le coefficient de corrélation définit l'intensité du lien entre les variables.
- Lorsque $r=0$, il n'y a pas de relation entre les variables
- Lorsque $r=1$, les 2 variables évoluent dans le même sens et la corrélation est dite positive.
- Si $r=-1$, les 2 variables évoluent dans les sens opposés et la corrélation est dite négative.

Régression

- la régression est un modèle de prédiction qui définit la relation de cause à effet entre les variables indépendantes et les variables dépendantes.
- on distingue les régressions simples avec 2 variables des régressions multiples où une variable est à expliquer à l'aide de plusieurs variables.
- Dans toute régression, on commence par construire le nuage de points. Si la corrélation n'est pas significative entre les variables, il n'est pas nécessaire de faire la régression.

Estimation des coefficients

- B_0 et b_1 pour les régressions simples, b_0 représente l'ordonnée à l'origine alors que b_1 est la pente de la droite.
- R^2 est le coefficient de détermination qui est la proportion de la variation totale qui est expliquée par le modèle. Ce coefficient doit être le plus élevé possible pour la validation du modèle.

Exercice pratique sur la corrélation/ régression

- L'exercice pratique a consisté à transférer vers R un jeu de données appelé Tp2. l' étape suivante consistait à effectuer les statistiques descriptives simples pour les variables X_1 , X_2 et X_3 .
- Faire des graphiques X_2 et X_3 en fonction de X_1
- Etablir et tester les corrélations (X_1, X_2) , (X_1, X_3) , (X_2, X_3)
- Etudier les régressions simples $(X_2 \sim X_1)$, $(X_2 \sim X_3)$
- Régressions multiples $\text{lm}(X_3 \sim X_1 + X_2)$

L'analyse en composantes principales

- Principes:
- Centrer les données ce qui revient à soustraire la moyenne à chaque donnée puis la diviser par l'écart type
- La diagonalisation consiste à construire la matrice constituée de vecteurs propres ou vecteurs Eigen
- L' étape suivante consiste à choisir les vecteurs propres et les valeurs propres.

L'analyse en composantes principales (2)

- L'étape suivante consiste à calculer les distance utilisées dans l'acp ce qui est la même chose que les distances euclidiennes entre deux points.
- A une distance donnée, les variables jouent le même rôle.
- Il faut user du bon sens en choisissant les axes car l'acp n'est pas une fin en soi.
- R possède 5 fonctions pour faire l'acp parmi lesquelles on peut citer princom(stat), pca(Factor MineR), acp(ama), dudi.pca(ade4) etc.

Exercice pratique

- La journée a été clôturée avec un exercice pratique sur l'analyse en composantes principales à faire à la maison.

COMPTE RENDU DU 11 SEPTEMBRE 2013

Rapporteurs du jour : MBONGO LOMBI Franklin (C2D Plantain) et ETAME François (C2D manioc)

Les travaux cette 3^e journée ont débuté à 9h02 min. L'ordre du jour comportait les points suivants :

- ❖ **La restitution des travaux de la journée du 10/09/2013**
- ❖ **La présentation et téléchargement du logiciel CURVE EXPERT**
- ❖ **Les techniques d'analyses factorielles multivariées**
 - **L'Analyse en Composante Principale (ACP suite)**
 - **L'analyse statistique bi variée avec des variables qualitatives (tableaux croisés)**
 - **L'Analyse Factorielle de correspondance (AFC)**

• **RESTITUTION DES TRAVAUX DE LA JOURNEE DU 10-09-2013**

La restitution des travaux de cette journée a été faite par les équipes des projets C2D Riz et C2D légumineuses représentées respectivement par Mr Bindzi Boniface et M. Mboussi à Messia. Les points essentiels de cette restitution étaient entre autre :

- **Rappels théoriques sur les notions de corrélation, de régression et de causalité**
- **Introduction sur les analyses multi variées : l'ACP (Analyse en Composante Principale)**
- **Exercice pratique sur les corrélations, régression et ACP (TP2)**

• **PRESENTATION ET TELECHARGEMENT DU LOGICIEL CURVE EXPERT**

Le formateur Dr CILAS. C a expliqué dans son exposé que les relations entre les variables ne sont pas toujours linéaire ; c'est-à-dire qu'on peut avoir un nuage de points dans une représentation graphique qui ne suit pas une droite. En d'autres termes la droite d'équation linéaire $Y=aX + b$ ne représente pas bien notre modèle.

Le logiciel **CURVE EXPERT** nous permet donc de pallier aux **problèmes de régressions non linéaires**. Ce logiciel est gratuit et peut être téléchargé sur internet sous la version *curve expert Basic 1.4*

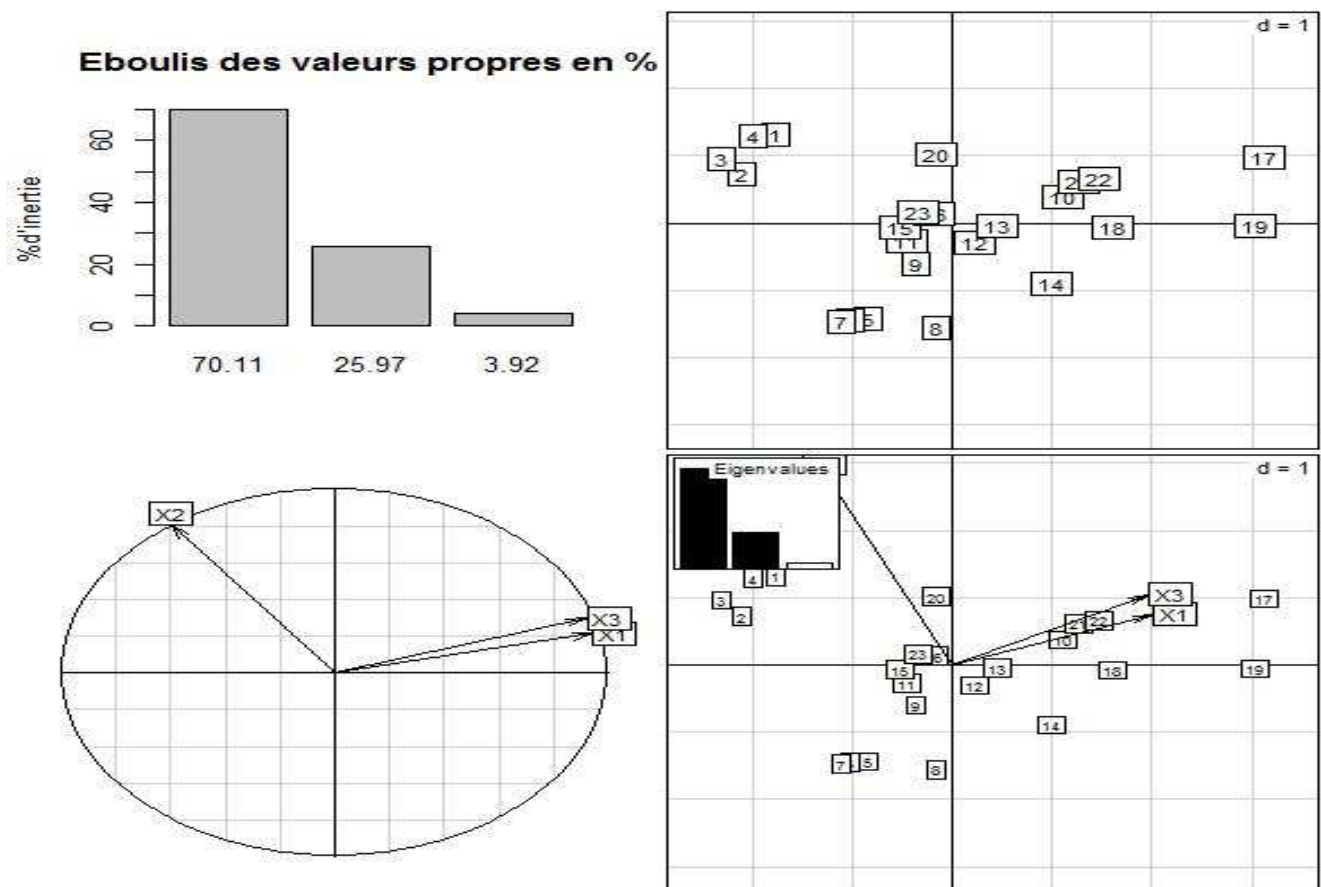
Il faut noter que parmi les modèles qui s'ajustent bien à nos données (R^2 élevé) non linéaire, on choisit celui qui a moins de paramètres et qui peut être facile à expliquer.

• **TECHNIQUES D'ANALYSES FACTORIELLES MULTI VARIEES**

- **TRAVAUX PRATIQUE SUR L'ACP (suite TP2)**

L'ACP en rappel est une technique **d'analyse descriptive** des données. On utilise l'ACP lorsque nous sommes en présence des variables **quantitatives** qui nécessitent au préalable une sorte de classification/catégorisation avant les analyses statistiques proprement dites (statistique inférentielle). La pratique de cette technique d'analyse par les participants a été faite à partir d'un jeu de données appelé **TP2** à la suite des exposés des formateurs

La syntaxe appropriée dans R a permis d'obtenir les résultats suivants :



Interprétation générale

Le 1er graphique (au-dessus à gauche) montre qu'environ 95 % de l'inertie est expliqué par les deux premières composantes principales qui sont en réalité des combinaisons des

Le 2^e graphique (en bas à gauche) nous montre que les variables X1 et X3 sont très corrélées, mais distantes de la variable X2.

Le 3^e graphique (en haut à droite) montre comment les individus sont repartis en classe ou groupe ; on peut par exemple noter que les individus 1,2,3 et 4 forment un groupe qui est différent des individus 17 et 19 qui semblent plus proches.

Le 4^e graphique (en bas à droite) est une association des graphiques 2 et 3

Il faut noter que les syntaxe [, -1] ou [2 :4] permettent de prendre en compte seulement les variables qui nous intéressent dans cette analyse des données

- **L'ANALYSE STATISTIQUE BIVARIEE AVEC DES VARIABLES QUALITATIVES**

Dans le cadre cette analyse, on s'intéresse à la **relation entre deux variables qualitatives**. On peut avoir soit :

- **Une relation de dépendance, c'est-à-dire on explique une variable en fonction d'une autre**
- **Une relation d'interdépendance**

Dans le cas de figure où on a deux variables qualitatives, **on choisit de faire une analyse des tableaux croisés** qui va nous permettre de mesurer la dépendance entre 2 variables. Le test de KHI-2 (mesure la distance) et le V de kramer (indice de mesure de la force de la relation entre deux variables) permettent de se prononcer sur la relation entre les deux variables. **Lorsque V=0 cela signifie qu'il n'y a aucune relation entre les variables. La relation de dépendance entre les variables est forte lorsque le KHI-2 et le V (V>0,7) sont élevés.**

La pratique de ce type d'analyse a été illustrée par un exemple de jeu de données qui porte sur **l'achat d'une marque**(variable dépendante) **et la situation matrimoniale** (variable indépendante) avec des effectifs

Lorsqu'on obtient après le t-test une p-value $< 0,05$, on peut conclure qu'il existe une relation de dépendance entre l'achat de la marque et la situation matrimoniale

- **L'ANALYSE FACTORIELLE DE CORRESPONDANCE (AFC)- Pratique-TP3**

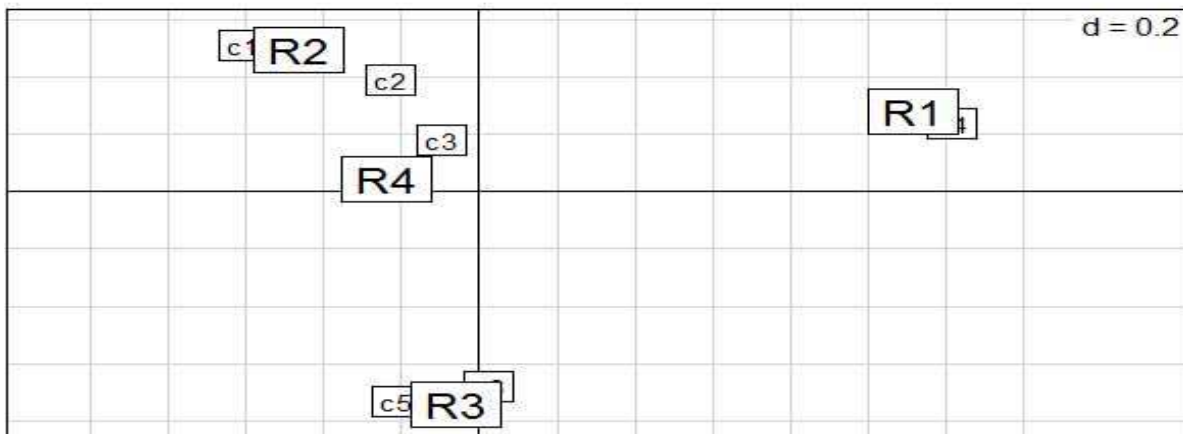
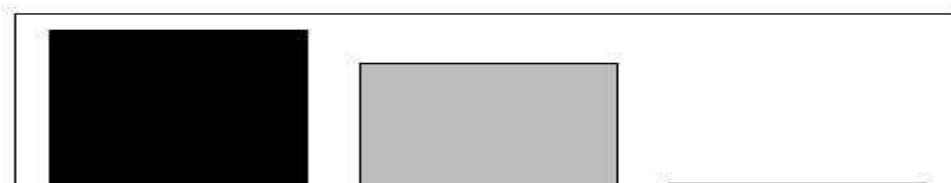
L'AFC plus que l'ACP cherche à mettre en correspondance deux ensembles de caractère, tout en générant le maximum d'information qu'on peut tirer à partir d'un tableau ou d'une matrice. Il s'agit d'une description des données **sous formes graphiques**

Dans une AFC, les lignes et les colonnes ont la même importance, ce qui amène à dire que l'AFC est une double ACP

Les fonctions suivantes sont nécessaires pour faire l'AFC : CA (FactorMiner) ; dudi.coa (ade4).

Un jeu de données qui fait appel à ce type d'analyse a été proposé par les formateurs et les résultats suivants ont été obtenus après avoir écrit le programme. La question principale de ce jeu de données est de savoir « **s'il existe une relation de dépendance entre les cultures et les régions** »

La syntaxe appropriée dans R a permis d'obtenir les résultats suivants :



Interprétation générale

On constate que les cultures c3, c2 et c1 semblent bien s'accommoder aux régions R2 et R4. La c4 est présente en R1 **Il s'agit en fait ici d'une description des données qui est un préalable à la statistique inférentielle**

On a obtenu par le test de KHI-2 une p-value $< 2.2e-16$ qui est très inférieure à 5% ; ce qui nous amène à conclure qu'il y'a une relation de dépendance entre les régions et les cultures. Cependant à ce niveau on ne peut pas encore dire quelles modalités des variables sont dépendantes les unes des autres

Les travaux de cette journée ont pris fin à 17h

COMPTE RENDU DU 12 SEPTEMBRE 2013

Rapporteurs : **Equipe Agroforesterie**

Ordre du jour

I-Rappel de la dernière séance

II-Analyses Multivariées (suite)

II.1. cas des Analyses Factorielles des Correspondances Multiples (AFCM)

II.2. Classification Ascendante Hiérarchique (CAH)

II.3. Analyses Discriminantes

II.4. Recommandations

IV- Introduction à la statistique spatiale

I- RAPPEL DE LA DERNIÈRE SÉANCE

□ ÉQUIPE C2D PLANTAIN ET MANIOC

II- ANALYSES MULTIVARIÉES

II.1. cas des Analyses Factorielles des Correspondances Multiples (AFCM)

- L'AFCM EST UNE SIMPLE EXTENSION DE L'ANALYSE FACTORIELLE DES CORRESPONDANCES. ELLE REVÊT UN INTÉRÊT EXEMPLE DE DOMAINE D'APPLICATION : L'ANALYSE DES DONNÉES ISSUES DES ENQUÊTES.

*PRÉALABLE

- CONSTRUCTION D'UN TABLEAU DISJONCTIF COMPLET (CODIFICATION PRÉSENCE ABSENCE {1,0})

*PRINCIPE À RESPECTER

- PAS D'EFFETIFS DE CLASSE TROP DÉSÉQUILIBRÉS
- DES NOMBRES DE CLASSES SEMBLABLES POUR TOUTES LES VARIABLES
- DES DÉCOUPAGES AYANT UNE SIGNIFICATION POUR LE CHERCHEUR

*APPLICATION

- LOGICIEL R PACKAGE ADE4 (FONCTION COA (CORRESPONDANCE ANALYSIS))
- **NB** : IL EST PRÉFÉRABLE DE PROCÉDER BIEN AVANT À UNE ANALYSE DESCRIPTIVE.

II.2. CLASSIFICATION ASCENDANTE HIÉRARCHIQUE

OBJECTIF : PRODUIRE DES GROUPEMENTS D'OBJETS OU D'INDIVIDUS DÉCRITS PAR DES VARIABLES OU DES CARACTÈRES EN CRÉANT DES PARTITIONS

*PRINCIPE DE REGROUPEMENT DES INDIVIDUS

- ON VA DU PLUS FIN AU PLUS PETIT, CHAQUE ÉLÉMENT D'UNE PARTITION EST INCLUS DANS UN ÉLÉMENT DE LA PARTITION SUIVANTE.

*PRÉALABLE (ETAPE DE L'ANALYSE)

- DISPOSER D'UN TABLEAU RECTANGULAIRE,
- MESURER LA DISTANCE ENTRE LES OBJETS : SIMILARITÉ OU DE DISSIMILARITÉ (KHI DEUX, EUCLIDIENNE ETC...)
- FORMER LES GROUPES LES PLUS HOMOGÈNES POSSIBLES EN FONCTION DES DISTANCES
- SEGMENTER (PERMET L'OBTENTION DU NOMBRE DE PARTITION)
- **APPLICATION** : MODULE HCLUST DU LOGICIEL R
- **NB** : POUR MIEUX INTERPRÉTER UNE ACP IL EST INDISPENSABLE DE LUI ASSOCIÉE UNE CAH
IL EST INDISPENSABLE DE DÉBUTER CETTE ANALYSE PAR UNE ACP

II.3. ANALYSE DISCRIMINANTE

- L'ANALYSE EST FAITE À PARTIR D'UN TABLEAU INDIVIDU/VARIABLE. L'OBJECTIF EST D'ESSAYER EN PRÉSENCE DES GROUPES D'AFFECTER UN INDIVIDU A UN GROUPE
- *RECHERCHER LES VARIABLES DISCRIMINANTES QUI SONT NON CORRÉLÉES ENTRE ELLES
- AFFECTER UNE NOUVELLE OBSERVATION À CHAQUE INDIVIDU EN FONCTION DES CLASSES
- **APPLICATION** : LOGICIEL R PACKAGE ADE4
- **NB** : L'ANALYSE DISCRIMINANTE EN PLUS DE SON ASPECT DESCRIPTIF COMPORTE UN PEU DE L'INFÉRENTIELLE

II.4. RECOMMANDATIONS

* QUESTIONS À SE POSER AVANT TOUTES ANALYSES

1. QUEL EST L'OBJECTIF PRINCIPAL DE LA RECHERCHE
2. COMBIEN DE VARIABLES DÉPENDANTES ET INDÉPENDANTES
3. QUELS TYPES D'ÉCHELLES DE MESURE
4. POSTULATS RENCONTRÉS OU NON

* OBJECTIFS DE LA RECHERCHE

1. ETABLIR DES RELATIONS ENTRE LES VARIABLES
2. RECHERCHER DES DIFFÉRENCES ENTRE LES GROUPES
3. PRÉDIRE L'APPARTENANCE À UN GROUPE
4. IDENTIFIER DES STRUCTURES

IV- INTRODUCTION À LA STATISTIQUE SPATIALE

IV.1. Rappels sur les notions de statistiques inférentielles

Objectifs: Décrire, classer, expliquer, comparer les groupes, prédire la valeur d'une variable à partir de la valeur d'autres variables

IV.2. Etude de la structure spatiale

X est continue dans l'espace	Variographie (Semivariogramme)
X est discrète	Fonction k de Ripley (processus ponctuel)
X est définie par des polygones	Moran et Geary (analyses des lattices)

COMPTE RENDU DU 13 SEPTEMBRE 2013

Rapporteurs : Eunice NDO (projet Horticulture) / Murielle EKALLE NYAME (projet aviculture)

L'ordre du jour de la journée s'articule comme suit :

- **Restitution de la journée du 12/09/13 par l'équipe d'agroforesterie**
- **Introduction à la statistique spatiale**
- **Analyse des données des participants**
- **Cérémonie de clôture**

I. Restitution de la journée du 12/09/13

Elle a débuté à 9h15. L'équipe d'agroforesterie a restitué les travaux de la journée du 12/09/13. Cette séance a été suivie par une série de présentations sur la statistique spatiale faite par Dr Cilas.

II. Introduction à la statistique spatiale

D'une manière générale on distingue 3 types de données spatiales :

- ***les données continues* : on fait appel à la géostatistique dans le but de faire de l'interpolation notamment pour la réalisation des cartes géologiques**
- ***les données lattices* : les sites sont des unités géographiques**
- ***les données ponctuelles* : pour étudier la répartition spatiale des individus d'une espèce sur un territoire.**

Dr Christian Cilas s'est attardé sur les données continues et les données ponctuelles.

a) Données continues :

Pour une interpolation on fait appel à deux méthodes

- **la méthode probabiliste qui étudie la structure spatiale (variogramme) et l'interpolation se fait par krigeage**
- **l'IDW (Inverse distance weighting) : qui utilise une matrice de poids non biaisée**

En général le krigeage est la méthode la plus appréciée car elle donne de meilleures images, des écarts type (des probabilités associées aux estimations) et évite les yeux.

b) modélisation spatiale

On distingue deux modèles de krigeage :

- **les modèles avec palier : sphérique, exponentiel, gaussien, cubique et sinus**

- **les modèles sans palier : puissance d'exposant \square et linéaire de pente \square**

c) Outils d'estimation et interpolation spatiale

L'objectif est d'essayer de modéliser la variation d'une variable quantitative en fonction de la distance entre les points.

➤ Le variogramme :

Son utilisation fait appel à l'ajustement à un modèle. Pour tester l'adéquation de ce modèle, soit on procède à une validation croisée (consistant à estimer tour à tour chaque observation par krigeage), soit on compare les performances de plusieurs modèles.

➤ krigeage

Dans ce chapitre le DrCilas a présenté les objectifs, les étapes et propriétés du Krigeage. A la fin de son propos quelques exemples ont été présentés. D'une manière générale on retient que le krigeage permet d'estimer la valeur d'un site en se basant sur un site voisin tout en tenant compte de sa structure spatiale.

Le logiciel R permet de faire du krigeage à l'aide des packages Gstat et Gorr

d) Etude de la structure et dynamique des communautés végétales

Trois parties ont été présentées à savoir :

- **Définition et principes généraux**
- **Processus ponctuel**
- **Quelques exemples**

D'une manière générale on retient que l'étude de la structure spatiale fait appel à la fonction $K(r)$ de Ripley et $L(r)$ qui permettent d'identifier trois types de structures : aléatoires, régulières ou agrégées.

Les packages utilisés avec le logiciel R sont ads et ar (études de biodiversité)

III. Analyse des données des participants

Les formateurs ont analysé et interprété les données de deux participants, l'un sur les nématodes et l'autre sur les variétés de haricot.

Après des échanges entre les participants et les formateurs, des fiches d'évaluation de nos formateurs ont été distribuées.

IV. Cérémonie de clôture de l'atelier

A 13 heures nous avons reçu une visite du Directeur régional du Cirad qui a tenu à féliciter les participants pour leur bonne tenue et à témoigner du plaisir qu'il a eu à nous recevoir dans les locaux de sa structure. Il a également émis le souhait que cette formation soit d'une grande utilité dans le cadre du projet C2D.

L'atelier s'est clôturé par une séance de remise des attestations de formation par les formateurs Dr Ndoumbè et Dr Cilas à 13h30.